Semi-Supervised Semantic Segmentation with Uncertainty-guided Self Cross Supervision

Yunyang Zhang[1], Zhiqiang Gong[1], Xiaoyu Zhao[1], Xiaohu Zheng[2], and Wen Yao[1]

[1] Defense Innovation Institute, Chinese Academy of Military Science, Beijing, China
zhangyunyang17@csu.ac.cn, wendy0782@126.com
[2] College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, China

Abstract. As a powerful way of realizing semi-supervised segmentation, the cross supervision method learns cross consistency based on independent ensemble models using abundant unlabeled images. In this work, we propose a novel cross supervision method, namely uncertainty-guided self cross supervision (USCS). To avoid multiplying the cost of computation resources caused by ensemble models, we first design a multi-input multi-output (MIMO) segmentation model which can generate multiple outputs with the shared model. The self cross supervision is imposed over the results from one MIMO model, heavily saving the cost of parameters and calculations. On the other hand, to further alleviate the large noise in pseudo labels caused by insufficient representation ability of the MIMO model, we employ uncertainty as guided information to encourage the model to focus on the high confident regions of pseudo labels and mitigate the effects of wrong pseudo labeling in self cross supervision, improving the performance of the segmentation model. Extensive experiments show that our method achieves state-of-the-art performance while saving 40.5% and 49.1% cost on parameters and calculations.

Keywords: Semi-Supervised Semantic Segmentation · Consistency Regularization · Multi-Input Multi-Output · Uncertainty.

1 Introduction

Semantic segmentation is a significant fundamental task in computer vision and has achieved great advances in recent years. Compared with other vision tasks, the labeling process for semantic segmentation is much more time and labor consuming. Generally, tens of thousands of samples with pixel-wise labels are essential to guarantee good performance for such a known data-hungry task. However, the high dependence of large amounts of labeled data for training would undoubtedly restrict the development of semantic segmentation. Semi-supervised semantic segmentation, employing limited labeled data as well as abundant unlabeled data for training segmentation models, is regarded as an effective approach to tackle this problem, and has achieved remarkable success for the task [17, 21, 28, 11, 10].

Advanced semi-supervised semantic segmentation methods are mainly based on consistent regularization. It is under the assumption that the prediction for the same object with different perturbations, such as data augmentation for input images [9, 17], noise interference for feature maps [28] and the perturbations from ensemble models [6, 14], should be consistent. Among these perturbations, the one through ensemble models usually provide better performance since it can learn the consistent correlation from each other adaptively. The earnings of a single model acquired from unlabeled images can be improved by cross supervision between models achieved by forcing consistency of the predictions.

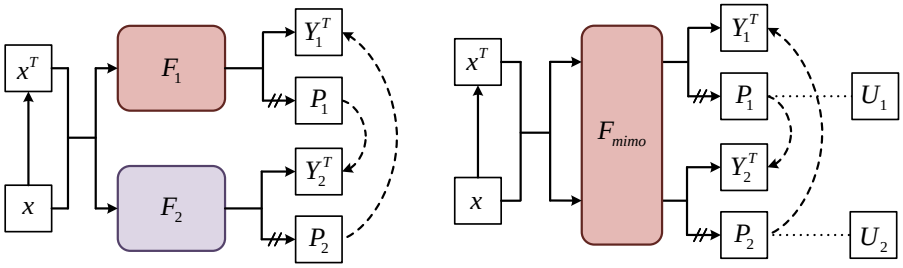

Fig. 1. Illustrating the architectures for (a) cross supervision and (b) our method uncertainty-guided self cross supervision. In our approach, x^T means the transformed image of x; Y_1^T and Y_2^T mean the predictions from a multi-input multi-output model F_{mimo}; P_1 and P_2 mean the pseudo labels for Y_1^T and Y_2^T; U_1 and U_2 mean the uncertainty of P_1 and P_2, respectively.

Despite impressive performance, the cost of time and memory for cross supervision is usually multiplicatively increased due to the parallel training of ensemble models with different model architectures or different initializations. To break this limitation, we propose a Self Cross Supervision method, which build only one model to obtain different views and significantly reduces the computation cost while achieving high performance.

Specifically, we impose cross supervision based on a multi-input multi-output (MIMO) model rather than multiple independent models. Commonly, one model hard to produce diversity. Thus we implicitly fit two subnetworks within single basic network utilizing the over-parameterization of neural network, achieveing diversity with single model. Through MIMO, multiple predictions can be obtained under a single forward pass, and then the purpose can be achieved almost "free" [13]. In our method, instead of ensemble models, cross supervision is realized by one MIMO model, which is called Self Cross Supervision.

Compared with multiple independent models, the performance of each subnetwork in one MIMO is compromised when the capacity of the MIMO is limited. The subnetwork with poor representation ability generate pseudo labels with large noise, further confusing the training process and making false propagation

from one subnetwork to others[46, 48, 47]. To suppress the noisy pseudo labels, we propose employing uncertainty guided the process of learning with wrong pseudo labels. Uncertainty is used to evaluate the quality of predictions without ground truth. Generally, regions with large uncertainty represent poor prediction and vice versa [1, 38]. For the task at hand, uncertainty can be used as the guided information to indicate the confidence of the pseudo labels of unlabelled samples and supervise the self cross supervision process by reducing the effects of wrong pseudo labeling, and such proposed method is called Uncertainty-guided Self Cross Supervision (USCS). The comparisons between cross supervision and our USCS is shown in Fig.1.

In conclusion, our contributions are:

1. We firstly propose a self cross supervision method with a multi-input multi-output (MIMO) model. Our method realizes cross supervision through enforcing the consistency between the outputs of MIMO, and greatly reduces the training cost of the model.
2. We propose uncertainty-guided learning for self cross supervision to improve the performance of the model, which uses the uncertainty information as the confidence of the pseudo labels and supervises the learning process by reducing the effects of wrong pseudo labeling.
3. Experiments demonstrate that our proposed model surpasses most of the current state-of-the-art methods. Moreover, compared with cross supervision, our method can achieve competitive performance while greatly reducing training costs.

2 Related work

2.1 Semantic segmentation

Semantic segmentation is a pixel-wise classification task, which marks each pixel of the image with the corresponding class. Most of the current semantic segmentation models are based on the encoder-decoder structure [2, 26, 30]. The encoder reduces the spatial resolution generating a high-level feature map, and the decoder gradually restores spatial dimension and details. Fully convolutional neural networks (FCN) [22] is the first encoder-decoder-based segmentation model. The subsequent works improve the context dependence by dilated convolutions [42, 4], maintaining high resolution [33, 37], pyramid pooling [44, 41], and self-attention mechanism [36]. DeepLabv3+ [5] is one of the state-of-the-art methods, which is employed as the segmentation model in this work.

2.2 Semi-supervised learning

Semi-supervised learning focuses on high performance using abundant unlabeled data under limited labeled data, so as to alleviate the training dependence on labels [19, 15, 45]. Most of the current semi-supervised learning methods are based

on empirical assumptions of the image itself, such as smoothness assumption, and low-density assumption [35].

Based on the smoothness assumption, prior works use the consistent regularization semi-supervised method, which encourage the model to predict the similar output for the perturbed input. This kind of works tries to minimize the difference between perturbed samples generated by data augmentations, e.g., Mean Teacher [34], VAT [24] and UDA [39]. As for the low-density assumption, the pseudo label based semi-supervised learning [19, 29, 39] is the representative method, which realizes the low-density separation by minimizing the conditional entropy of class probability for the unlabeled data. In order to utilize the merits of different assumptions, prior works also propose effective methods based on both or more. Among these methods, joint learning with the pseudo label and consistent regularization is a successful one and has achieved impressive performance, such as MixMatch [3], FixMatch [32] and DivideMix [20]. Our approach utilizes consistent regularization and the pseudo label to construct semi-supervised learning.

2.3 Semi-supervised semantic segmentation

As a dense prediction task, semantic segmentation is laborious and time-consuming in manual annotations. Therefore, using unlabeled images to improve model performance is an effective way for cost reduction. Most of the semi-supervised semantic segmentation approaches are based on the consistent regularization [27, 9, 49, 17].For example, PseudoSeg [49] enforces the consistency of the predictions with weak and strong data augmentations, similar to FixMatch [32]. CAC [17] utilizes contextual information to maintain the consistency between features of the same identity under different environments. CCT [28] maintains the agreement between the predictions from the features with various perturbations. GCT [14] and CPS [6] adopt different model structures or model initializations to generate the perturbations of predictions and achieve state-of-the-art performance. However, the training cost of time and memory for ensemble models is expensive in GCT and CPS. Different from prior works, our approach enforces the consistency of predictions from a multi-input multi-output network and greatly reduces the training costs.

3 Method

In the following sections, we first introduce the overview of USCS in Sec. 3.1. The self cross supervision with MIMO model is proposed in Sec. 3.2. To ameliorate pseudo label quality, we propose the uncertainty-guided learning in Sec. 3.3.

As a common semi-supervised learning task, a dataset $\mathcal{X}$ consisting of labeled images $\mathcal{X}_l$ with labels $\mathcal{Y}$ and unlabeled images $\mathcal{X}_{ul}$ is employed to train a segmentation network. In our USCS, we extra applied transformation T on unlabeled images $\mathcal{X}_{ul}$ got the transformed images $\mathcal{X}_{ul}^T = T(\mathcal{X}_{ul})$. Both unlabeled images $\mathcal{X}_{ul}$ and transformed images $\mathcal{X}_{ul}^T$ are employed to construct self cross supervision.

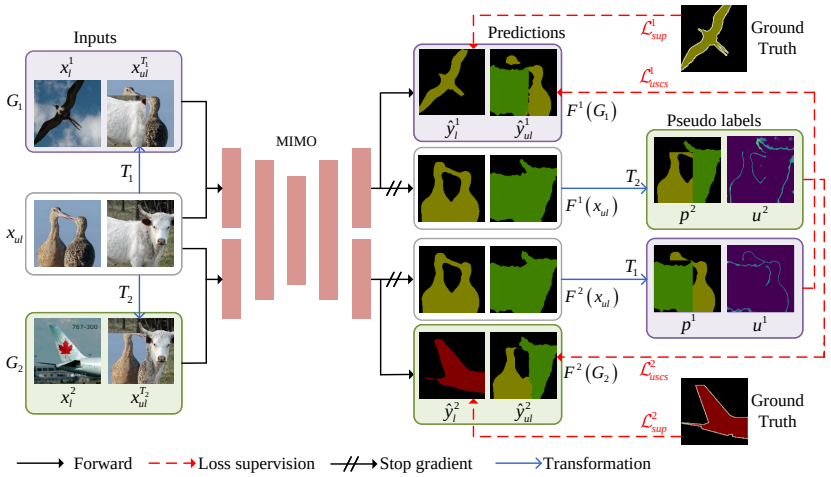

Fig. 2. The USCS Framework. We aim to maintain the consistency between the predictions from a multi-input multi-output (MIMO) model. Since MIMO accepts two different group images, we adopted transformation consistency to realize the purpose.

3.1 Overview of USCS

The USCS framework is shown in Fig. 2. In contrast to the general cross supervision method using several independent models, we instead employ a multi-input multi-output (MIMO) model. Specifically, the MIMO model F has two input and output branches which can be seen as the subnetworks with shared parameters, accepting two groups independently sampled data $G_k\,(k \in \{1,2\})$ and output corresponding segmentation results. In USCS, each group data is denoted as $G_k = \left\{x_l^k, x_{ul}^{T_k}\right\}$ $(x_l^k \in \mathcal{X}_l, x_{ul}^{T_k} \in \mathcal{X}_{ul}^T)$. For $k \in \{1,2\}$, x_l^1 and x_l^2 are the labeled images with different batch sampling order, $x_{ul}^{T_1}$ and $x_{ul}^{T_2}$ are the transformed images with distinct transformation T_k.

Given an image $x^k \in G_k$, the MIMO model F first predicts $\hat{y}^k = \{\hat{y}_l^k, \hat{y}_{ul}^k\}$, where $\hat{y}_l^k = F(x_l^k)$ and $\hat{y}_{ul}^k = F(x_{ul}^{T_k})$. As common semantic segmentation models, the prediction $\hat{y}_l^k$ is supervised by its corresponding ground-truth $y \in \mathcal{Y}$ as:

$$\mathcal{L}_{sup}^k\left(x_l^k, y\right) = \frac{1}{|\Omega|}\sum_{i\in\Omega}\ell_{ce}(\hat{y}_l^k(i), y(i)), \tag{1}$$

where $\ell_{ce}(*)$ is the standard Cross Entropy loss, and Ω is the region of image with size $H \times W$.

To explore the unlabeled images, we repeat the original unlabeled images x_{ul} twice, the MIMO model F makes two groups independent predictions $F^1(x_{ul})$ and $F^2(x_{ul})$ on the same images x_{ul} as shown at the bottom of Fig. 2. Then the same transformation T_1 and T_2 are respectively performing on $F^2(x_{ul})$ and $F^1(x_{ul})$, obtaining $p^1 = T_1(F^2(x_{ul}))$, $p^2 = T_2(F^1(x_{ul}))$. Besides, the uncertainties u^1 and u^2 are estimated for two transformed predictions p^1 and p^2,

respectively. Then, p^1 guided by the uncertainty u^1 is regarded as the pseudo labels of $x_{ul}^{T_1}$ to supervise $\hat{y}_{ul}^1$. Similarly, the same operation is used to supervise $\hat{y}_{ul}^2$ based on p^2 and u^2. We call the above process uncertainty-guided self cross supervision. The constraint $\mathcal{L}_{uscs}$ and more details are described in Sec. 3.2 and Sec. 3.3.

Finally, our method for the training of MIMO model F joint the two constraints on both the labeled and unlabeled images which can be written as:

$$\mathcal{L}\left(\mathcal{X},\mathcal{Y}\right)=\sum_{k=1,2}\left(\frac{1}{|\mathcal{X}_l|}\sum_{x_l^k\in\mathcal{X}_l}\mathcal{L}_{sup}^k\left(x_l^k,y\right)+\frac{1}{|\mathcal{X}_{ul}|}\sum_{x_{ul}\in\mathcal{X}_{ul}}\lambda\mathcal{L}_{uscs}^k(x_{ul})\right),\tag{2}$$

where λ is the trade-off weight to balance the USCS constraint.

3.2 Self Cross Supervision with MIMO Model

The proposed self cross supervision is implemented over the MIMO model. Before presenting self cross supervision, the MIMO model used in USCS is firstly introduced. Based on the fact that neural networks are heavily overparameterized models [13], we can train a MIMO model containing multiple independent subnetworks and acquire multiple predictions of one input under a single forward pass of the model. Different from the single neural network archtiecture, the MIMO model replaces the single input layer by N input layers, which can receive N datapoint as inputs. And N output layers are added to make N predictions based on the feature before output layers. Compared with a single model, the MIMO model obtains the performance of ensembling with the cost of only a few increased parameters and calculations.

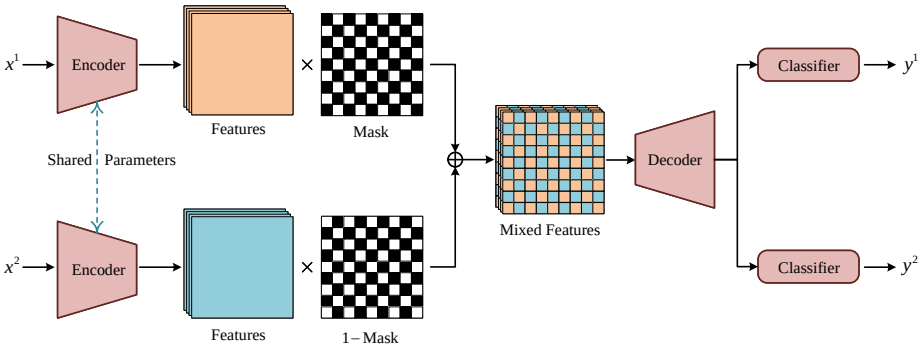

Fig. 3. The structure of MIMO segmentation model. The features after the encoder are fused by the grid mix.

In USCS, we construct a MIMO model with two inputs and outputs, whose structure is shown in Fig. 3. For better extract object features, the entire encoder part is utilized as the input layer of the model (the original MIMO model

employs the first convolutions layers of the model as the input layer). However, two independent encoders (the input layer) increase the model parameters and computation. We share the parameters of two encoders to avoid this problem. The features of two inputs extracted by the encoder must be fused before entering the decoder. To effective combines inputs into a shared representation, the grid mix is adopted to replace the original summing method [13] in MIMO as:

$$\mathcal{M}_{gridmix}(f_1, f_2) = \mathbb{1}_{\mathcal{M}} \odot f_1 + (\mathbb{1} - \mathbb{1}_{\mathcal{M}}) \odot f_2, \tag{3}$$

where f_1 and f_2 are the features of two inputs, respectively; $\mathbb{1}_{\mathcal{M}}$ is a binary grid mask with grid size g.

The self cross supervision enforces two predictions of MIMO learn from each other. The output y^1 is considered the pseudo label to supervise the output y^2, vice versa. As mentioned previously, two inputs of MIMO are different, while the self cross supervision is feasible only when the inputs are the same. We overcome this issue by introducing the transformation consistency regularization [25], which assumes that the prediction $F(T(x))$ of the transformed image $T(x)$ must be equal to the transformed prediction $T(F(x))$ of the original image x.

As shown in Fig. 2, the MIMO model F predicts two transformed unlabeled images $x_{ul}^{T_1}$ and $x_{ul}^{T_2}$, obtaining $\hat{y}_{ul}^1$ and $\hat{y}_{ul}^2$. Self cross supervision expects two outputs of the MIMO model to supervise each other. However, the semantics of the outputs $\hat{y}_{ul}^1$ and $\hat{y}_{ul}^2$ are different. To achieve the self cross supervision, we input the original unlabeled image x_{ul} to the MIMO model, getting two individual predictions $F^1(x_{ul})$ and $F^2(x_{ul})$ without gradient. We further obtain two transformed predictions $p^1 = T_1(F^2(x_{ul}))$ and $p^2 = T_2(F^1(x_{ul}))$ by performing the transformation T_1 and T_2, respectively. The transformed predictions p^1 should have the similar semantics with $\hat{y}_{ul}^1$, thus we regard p^1 as the pseudo label of $\hat{y}_{ul}^1$. Similarly, the transformed prediction p^2 is considered as the pseudo label to supervise $\hat{y}_{ul}^2$.

Through the above process, the MIMO model F can realize cross supervision by itself. The self cross supervision constraint on unlabeled data is defined as:

$$\mathcal{L}_{scs}^k(x_{ul}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} \ell_{ce}(\hat{y}_{ul}^k(i), p^k(i)). \tag{4}$$

3.3 Uncertainty-guided Learning

The pseudo label obtained from the prediction exists noise, especially when the capacity of subnetworks in MIMO is limited. The poor model representation leads to plenty of inaccurate pseudo labels. The noisy pseudo label will mislead the model and interfere with the optimization direction in self cross supervision. In addition, the noise caused by one model is likely to propagate to another model through self cross supervision, resulting in the accumulation and propagation of errors and hindering the performance. It is necessary to filter the pseudo label with inferior quality to improve the overall performance of the model.

Uncertainty estimation is an effective method to evaluate noise in prediction [18]. Noise often exists in regions with large uncertainties. Fig. 4 shows

the uncertainty visualization. Based on this observation, we propose to employ uncertainty to guide the pseudo label with noise in cross supervision. Firstly, we estimate the uncertainty of pseudo label through the Shannon Entropy [31], which is defined as:

$$U = -\sum_{c=1}^{C} p(c) \log p(c), \tag{5}$$

where C is the softmax predicted class related to the category of the dataset,

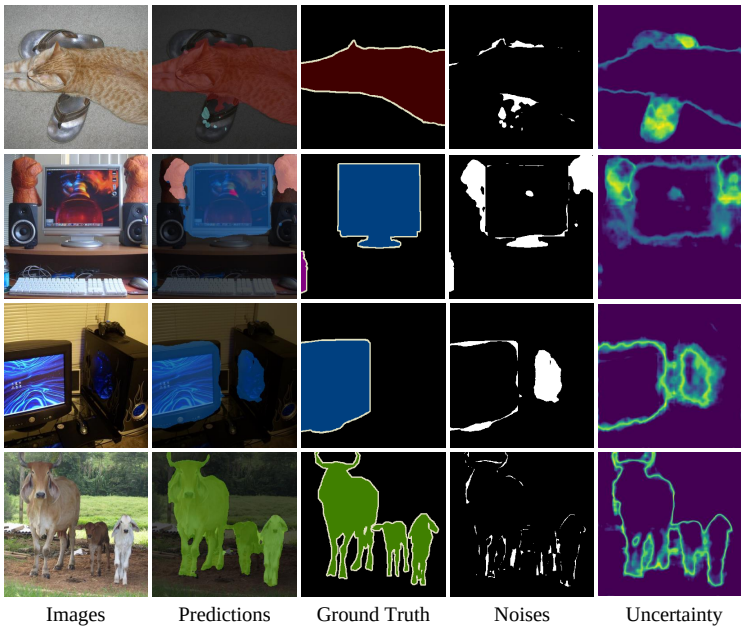

Fig. 4. Uncertainty visualization. Highly bright regions represent large uncertainties in the uncertainty map.

p is the softmax predicted vector with C. We normalize U into range $(0, 1)$, and set $\widehat{U} = 1 - U$. Then, the pseudo label can be divided into confident and uncertain regions by setting a threshold γ. We fully receive the pixels in the confident region, which are regarded as the true label. As for the uncertain regions, we assign low loss weights to high uncertain pixels. Thus the model can also learn from the pixels in the uncertain regions, which avoids the loss of useful information. We define the uncertainty weight mask as:

$$W = \begin{cases} 1 & \widehat{U} \geq \gamma \\ \widehat{U}/\gamma & \widehat{U} < \gamma \end{cases} \tag{6}$$

In the end, we multiply the weight mask W to the self cross supervision constraint and rewrite the Eq. 4, getting the uncertainty-guided self cross supervision constraint:

$$\mathcal{L}_{uscs}^{k}(x_{ul}) = \frac{1}{|\Omega| \, \|W\|_{1,1}} \sum_{i \in \Omega} W(i) \cdot \ell_{ce}(\hat{y}_{ul}^{k}(i), p^{k}(i)), \tag{7}$$

where $\|W\|_{1,1}$ means the $L_{p,q}$ norm of matrix W.

4 Expermients

4.1 Experimental Setup

Datasets. PASCAL VOC 2012 [8] is the most prevalent benchmark for semi-supervised semantic segmentation with 20 object classes and one background class. The standard dataset contains 1464 images for training, 1449 for validation, and 1456 for testing. Following previous works [6], we adopt the augmented set provided from SBD [12] as our entire training set, which contains 10582 images.

Implementation details. The results are obtained by training the MIMO model, modified on the basis of Deeplabv3+ [5]. We regard the backbone of the segmentation model as the encoder, whose weights are initialized with the pre-trained model on ImageNet [7]. The other components except the final classifier are considerd as the decoder which are initialized randomly.

Following the previous works [6], we utilize "poly" learning rate decay policy where the base learning rate is scaled by $(1 - iter/max\_iter)^{0.9}$. Mini-batch SGD optimizer is adopted with the momentum and weight decay set to 0.9 and 10^{-4} respectively. During the training, images are randomly cropped to 320×320, random horizontal flipping with a probability of 0.5, and random scaling with a ratio from 0.5 to 2.0 are adopted as data augmentation. We train PASCAL VOC 2012 for 3×10^4 iters with batch size set to 16 for both labeled and unlabeled images. The base learning rates are 0.01 for backbone parameters and 0.001 for others. The trade-off weight λ is set to 1 after adjustment.

Besides, we found that the MIMO model based on Deeplabv3+ cannot accommodate two independent subnetworks due to the limited capacity. Thus, we relax independence same as [13] by sampling two same inputs from the training set with probability ρ, i.e., the input x_2 of the MIMO model is set to be equal to x_1 with probability ρ. During the training, we employ CutMix [43] as transformation, same as [6]. We average two outputs of the MIMO model to generate the final results for evaluation.

Evaluation. We use the mean Intersection-over-Union (mIoU) as the evaluation metric as a common practice. To evaluate training time and memory cost reduction in USCS, Multiply–Accumulate Operations (MACs) and the number of parameters are adopted as the metric. Besides, we employ the non overlap ratio for the outputs of the MIMO model as metric to measure the diversity of subnetworks. The low non overlap ratio means poor diversity.

4.2 Results

In this section, we report the results compared with supervised baselines and other SOTA methods in different partition protocols, i.e., the full training set is split with 1/16, 1/8, 1/4, and 1/2 ratios for labeled images and the remainder as unlabeled images.

Improvements over Supervised Baselines. Fig. 5 illustrates the improvements of our approach compared with full supervised learning (trained with the same partition protocol). Specifically, our method outperforms the supervised baseline by 5.70%, 4.23%, 2.63%, and 1.30% under 1/16, 1/8, 1/4, and 1/2 partition protocols separately with Resnet-50. On the other settings, the gains obtained by our approach are also stably: 5.20%,3.32%, 2.03%, and 1.50% under 1/16, 1/8, 1/4, and 1/2 partition protocols separately with Resnet-101.

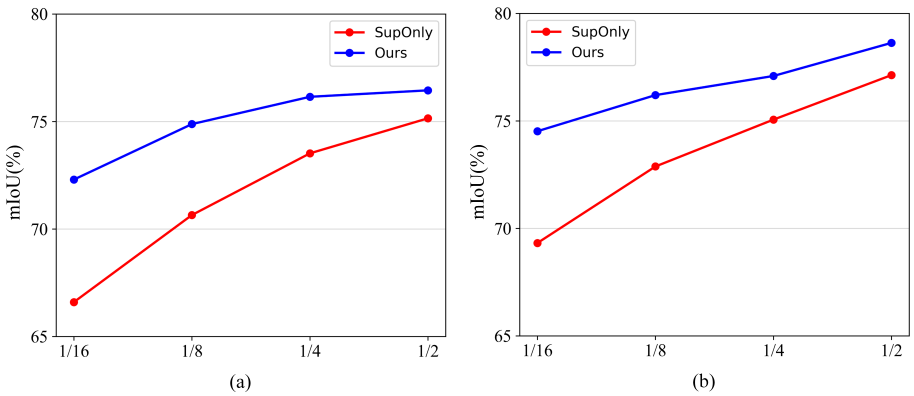

Fig. 5. Improvements over the supervised baseline on PACAL VOC 2012 with (a) Resnet-50 and (b) Resnet-101

Comparison with SOTA. The results compared with other semi-supervised approaches are shown in Tab. 1. Our method performs better than most methods under different partition protocols with Resnet-50 and Resnet-101 as backbones. Compared with CAC [17], our approach improves by 2.2%, 2.48%, 2.15% under 1/16, 1/8, and 1/4 partition protocols separately with Resnet-50. Compared with CPS [6], the advantage of our method is a great reduction in the number of parameters and calculations as shown in Tab. 2. We acquired 40.5% and 49% economization on MACs and parameters with Resnet-50, which signify the cost decrease of training time and memory. Besides, our method only needs twice forward pass, while CPS needs four times. As for accuracy, our method achieves around 1% improvement in all cases with Resnet-50.

Table 1. Comparison with SOTA on PASCAL VOC 2012. All the approachs are based on Deeplabv3+. The $*$ indicates the approaches re-implemented by [6]. Best results are in bold; suboptimal results are in italics.

Methods	Network	1/16	1/8	1/4	1/2
MT$*$[34]		66.77	70.78	73.22	75.41
CutMix-Seg$*$[9]		68.90	70.70	72.46	74.49
CCT$*$[28]		65.22	70.87	73.43	74.75
GCT$*$[14]	Deeplabv3+	64.05	70.47	73.45	75.20
CAC[17]	Resnet50	70.10	72.40	74.00	-
ECS[23]		-	70.20	72.60	-
CPS[6]		*71.98*	*73.67*	*74.90*	*76.15*
Ours		**72.30**	**74.88**	**76.15**	**76.45**
MT$*$[34]		70.59	73.20	76.62	77.61
CutMix-Seg$*$[9]		72.56	72.69	74.25	75.89
CCT$*$[28]		67.94	73.00	76.17	77.56
GCT$*$[14]	Deeplabv3+	69.77	73.30	75.25	77.14
CAC[17]	Resnet101	72.40	74.60	76.30	-
ELN[16]		-	75.10	76.58	-
ST++[40]		*74.50*	**76.30**	*76.60*	-
Ours		**74.52**	*76.20*	**77.09**	**78.63**

4.3 Ablation Study

This section conducts the ablation study to exhibit the roles of self cross supervision (SCS) and uncertainty-guided learning (UL) in our method. Besides, the influences of uncertainty threshold γ, feature fusion methods, and input repetition probability ρ are reported, respectively. All the experiments are run based on 1/8 partition protocols on PASCAL VOC 2012.

Table 2. Training cost comparison with CPS [6] and SupOnly in the backbone of Resnet-50 and Resnet-101.

Methods	Resnet-50		Resnet-101		Forwardings
	MACs(G)↓	Params(M)↓	MACs(G)↓	Params(M)↓	
SupOnly	23.84	39.78	31.45	58.77	1
CPS	95.36	79.56	125.80	117.54	4
Ours	56.74	40.49	71.94	59.48	2

Uncertainty guided self cross supervision. The contribution of self cross supervision and uncertainty-guided learning are shown in Tab. 3. It is important to note that we adopt the result of CPS with CutMix augmentation[6] as

a baseline to ensure fairness. We report a slight decline in performance after replacing CPS with SCS. While, the improvements yielded by UL are 1.23% with the Resnet-50. We can see that SCS heavily reduces training costs of time and memory, and UL improves the performance without extra cost.

Table 3. Ablation study of different components under 1/8 partition protocols on PASCAL VOC 2012.

CPS	SCS	UL	Deeplabv3+ with Resnet-50		
			mIoU(%)↑	MACs(G)↓	Params(M)↓
✓			73.67	95.36	79.56
	✓		73.65	56.74	40.49
	✓	✓	**74.88**	**56.74**	**40.49**

Uncertainty threshold γ. We investigate the influence of threshold γ used to control the uncertain weight mask as shown in Equation. 6. The results in Fig. 6(a) show that: with the increase of γ, the model reduces the weight of learning for noisy pixels in pseudo label and performs best when $\gamma = 0.5$. When the continuous increase of γ, the performance degrades due to the model tends to regard all pixels in pseudo label as noise, reducing the weight of confident pixels in pseudo label. We visualize the effect of threshold γ to uncertainty in Fig. 7.

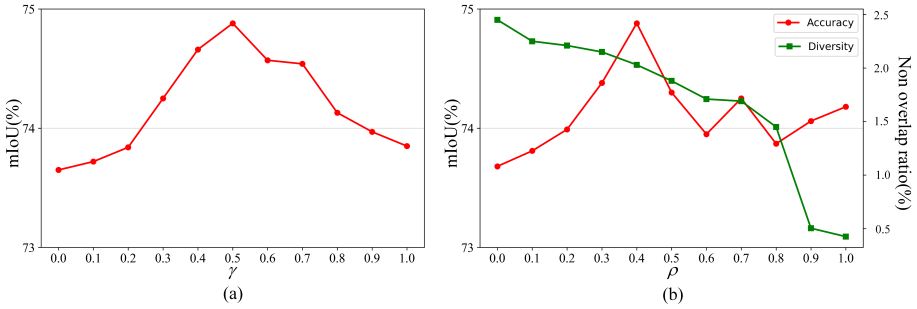

Fig. 6. The ablation study on (a) uncertainty threshold γ and (b) input repetition probability ρ.

Input repetition probability ρ. We show the influence of probability ρ on both accuracy and diversity in Tab. 6. When $\rho = 0$, the training images are

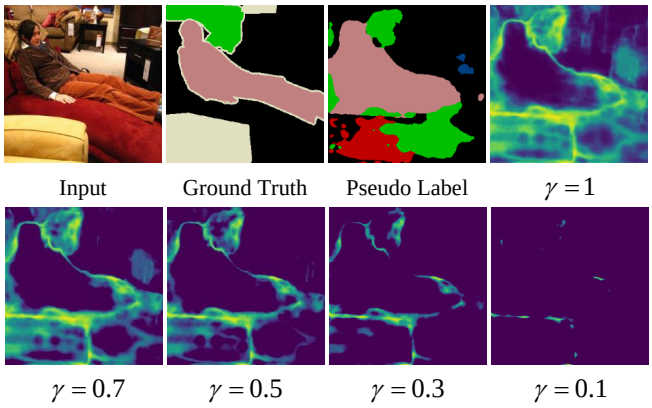

Fig. 7. Visual comparison with different uncertainty threshold γ.

sampled independently for both subnetworks, and the MIMO model acquired great diversity but poor accuracy as it can not contain two independent subnetworks. As ρ grew, the diversity of the MIMO model gradually decayed, the independence of the subnetwork is relaxed to release the limited model capacity. The performance reaches the peak at $\rho = 0.4$, where get a trade-off between the diversity and the capacity of the MIMO model.

Table 4. The affects of feature fusion methods on mIoU(%) and non overlap ratio(%).

Fusion methods	Grid mix				Summing
	1	3	5	7	
mIoU(%)	74.88	74.10	73.45	73.64	73.32
Non overlap ratio(%)	2.03	2.53	1.64	1.40	0.94

Feature fusion. We show the influence of feature fusion methods, summing and grid mix, on both accuracy and diversity in Tab. 4. The block size g of the grid mix is set as 1, 3, 5, and 7. We can see that the grid mix surpasses the summing feature fusion method on both mIoU scores and non overlap ratios. The accuracy of the MIMO model decreases as g increases, while the diversity reaches the top at $g = 3$. We use $g = 1$ in our method for all the experiments.

Qualitative Results. Fig. 8 visualizes some segmentation results on PASCAL VOC 2012. The supervised results display the bad accuracy caused by the limited labeled training samples. For example, in the 2-nd row, the supervised baseline

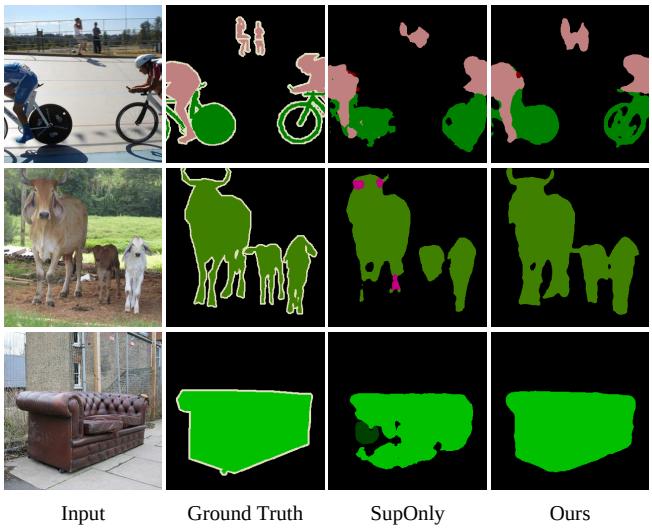

Fig. 8. Qualitative results from Pascal VOC 2012.

mislabels the cow as the horse in many pixels. While our method successfully corrected the wrong annotation. Besides, the segmentation labeled by our method is more exquisite than the supervised-only method.

5 Conclusions

In this paper, we propose a new cross supervision based semi-supervised semantic segmentation approach, Uncertainty-guided Self Cross Supervision (USCS). Our method achieves self cross supervision by imposing the consistency between the subnetworks of a multi-input multi-out (MIMO) model, avoiding high computations from ensemble training. Limited by the model capacity, the subnetwork representation ability of MIMO is poor, resulting in large pseudo label noise. In order to alleviate the problem of noise accumulation and propagation in the pseudo label, we proposed uncertainty-guided learning, utilizing the uncertainty as guided information to reduce the effects of wrong pseudo labeling. Experiments show our approach dramatically reduces training costs and achieves powerful competitive performance.

Acknowledgements This work was supported by the Natural Science Foundation of China under Grant 62001502.

References

1. Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra

Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.
2. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
3. David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.
4. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
5. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
6. Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.
7. MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.
8. Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.
9. Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. 2019.
10. Zhiqiang Gong, Ping Zhong, and Weidong Hu. Statistical loss and analysis for deep learning in hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):322–333, 2020.
11. Zhiqiang Gong, Ping Zhong, Yang Yu, Weidong Hu, and Shutao Li. A cnn with multiscale convolution and diversified metric for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3599–3618, 2019.
12. Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.
13. Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*, 2020.
14. Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, pages 429–445. Springer, 2020.
15. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
16. Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9957–9967, 2022.

17. Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021.
18. Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
19. Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013.
20. Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.
21. Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. *arXiv preprint arXiv:2111.12903*, 2021.
22. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
23. Robert Mendel, Luis Antonio de Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *European Conference on Computer Vision*, pages 141–157. Springer, 2020.
24. Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.
25. Aamir Mustafa and Rafał K Mantiuk. Transformation consistency regularization–a semi-supervised paradigm for image-to-image translation. In *European Conference on Computer Vision*, pages 599–615. Springer, 2020.
26. Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.
27. Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.
28. Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.
29. Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.
30. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
31. Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.
32. Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

33. Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.
34. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.
35. Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.
36. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
37. Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.
38. Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9092–9101, 2021.
39. Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.
40. Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022.
41. Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.
42. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
43. Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
44. Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
45. Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.
46. Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R Manmatha, Mu Li, and Alexander J Smola. Improving semantic segmentation via efficient self-training. *IEEE transactions on pattern analysis and machine intelligence*, 2021.
47. Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.
48. Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.
49. Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.